

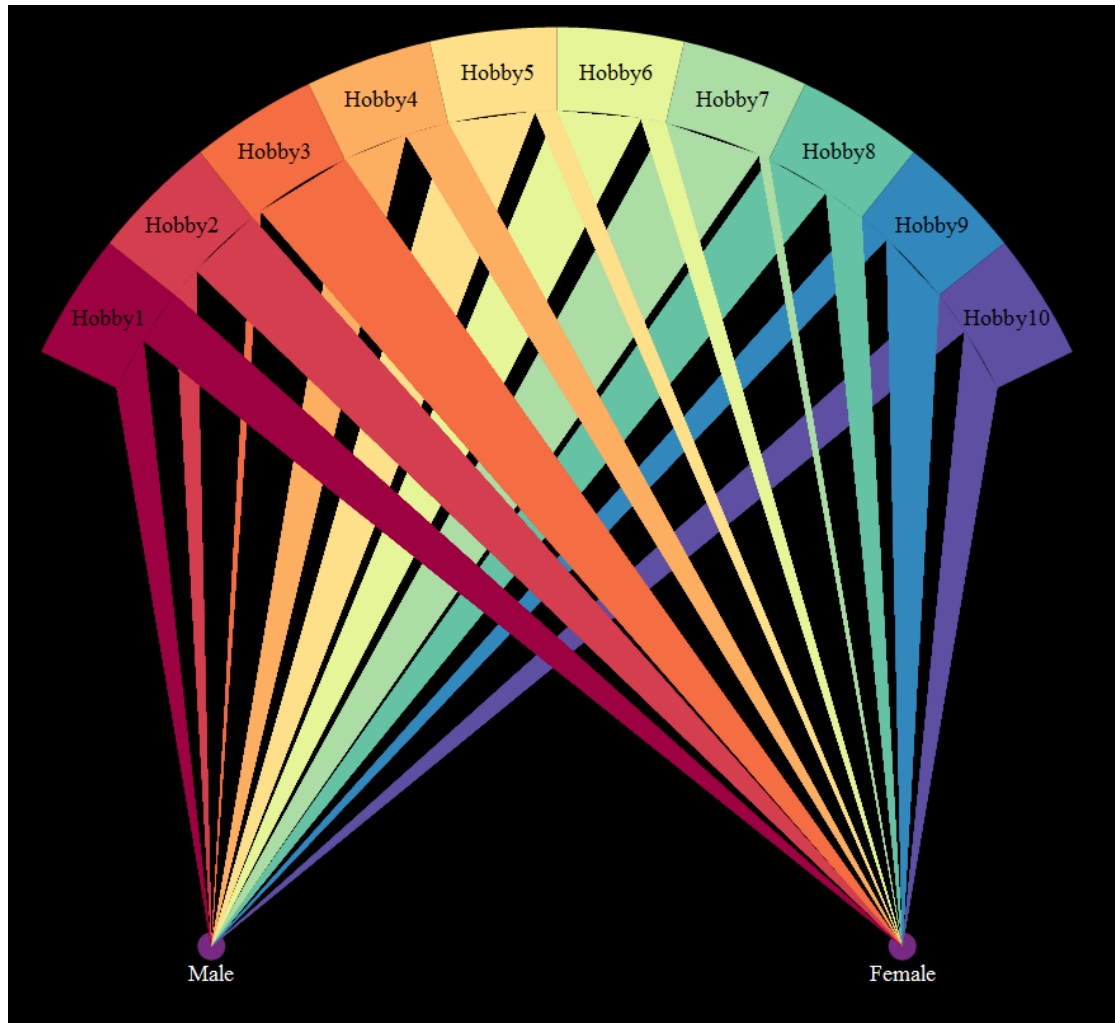
Weekly Report

2012.11.12-2012.11.18

黄芯芯

本周工作：

1. 这周考试周没有考试，所以都在淘宝上班。完成了淘宝那边布置的一个可视化小任务。
如下图所示：



2. 淘宝标签项目：

这周对淘宝标签项目有一个新想法，就是结合分析交易数据常用的模型“Market Basket Analysis”，然后通过对关联项进行计算和可视化，进而分析用户特征。

1) Market Basket Analysis （购物篮分析）

购物篮分析的过程是通过发现顾客放入其购物篮中不同商品之间联系，分析顾客的购买习惯。通过了解哪些商品频繁地被顾客同时购买，这种关联的发现可以帮助零售商制定营销策略。例如，在同一次去超级市场，如果顾客购买牛奶，他也购买面包（和什么类型的面包）的可能性有多大？通过帮助零售商有选择地经销和安排货架，这种信息可以引导销售。例如，将牛奶和面包尽可能放近一些，可以进一步刺激一次去商店同时购买这些商品。

2) 关联规则

反映商品频繁关联或同时购买的模式可以用关联规则来表示，如下规则表示买计算机也同时趋于购买杀毒软件的关联规则：

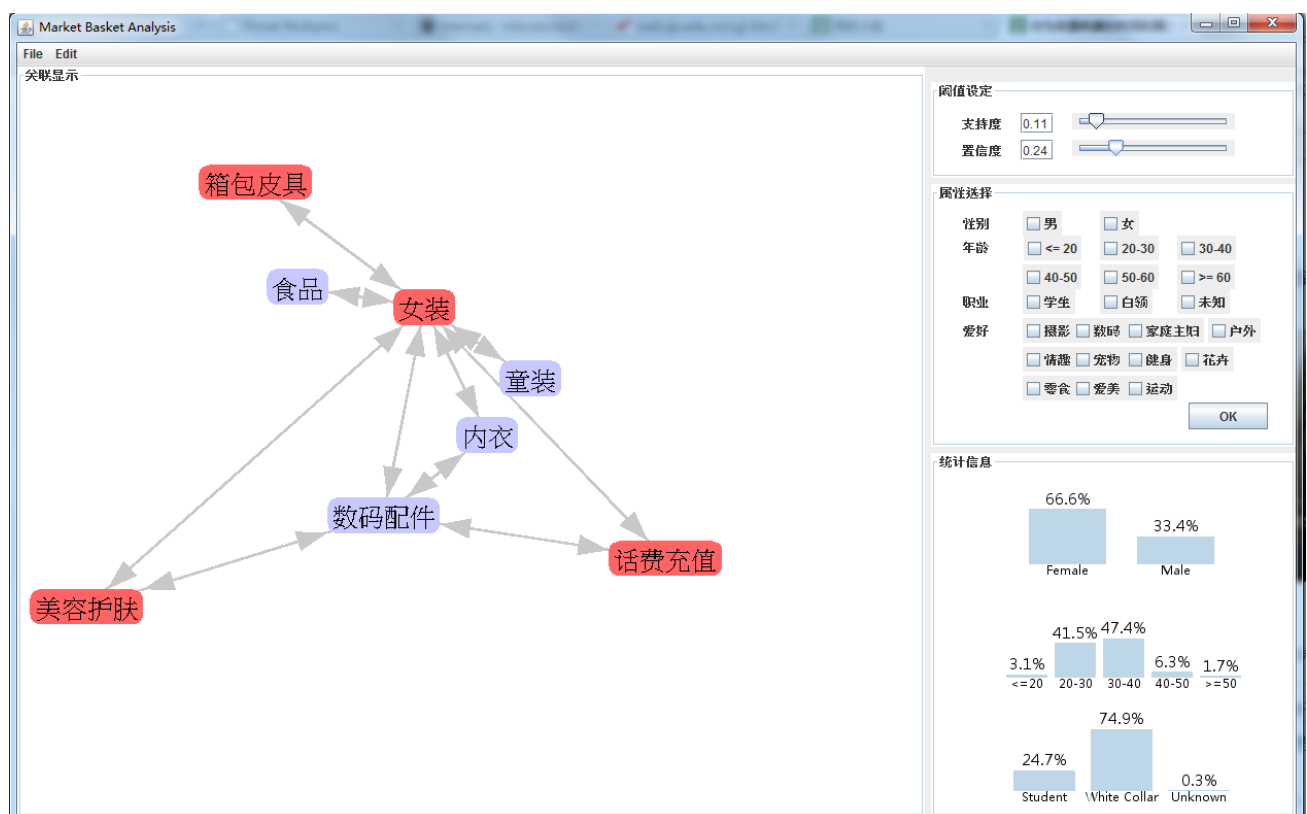
Computer → Anti-virus Software
[support = 2%, confidence = 60%]

其中，support 称为支持度，表示在全部交易中，computer 和 anti-virus software 同时出现在一笔交易中所占的百分比，反映了关联规则的有用性。confidence 称为置信度，表示在所有购买了 computer 的交易中，同时也购买了 anti-virus software 的所占的百分比，反映了关联规则的确定性。

3) 用于淘宝数据

其实有点参考马老师 click stream 那篇文章的思路：将用户行为分析结果可视化，然后交互选择一些行为模式，查看具有这些行为模式的用户的各种属性，再进行分析。但是如何能把可视化做得更好是还要思考的问题，比如说布局上、分析模型上等等。

将关联规则计算用于淘宝数据的思路是：计算每两个类目之间的关联规则，得到一个有向图，每个节点代表一个类目，节点之间的有向边表示两个类目之间的关联规则，然后用户交互选择某些感兴趣的关联规则，选择之后，符合这些关联规则的用户统计数据会通过简单的柱状图显示。如下图所示：



- 左边的面板显示了经过关联规则计算后，用力引导模型将关联规则可视化，只有大于最小支持度和置信度的关联规则才会被显示。关联规则具有方向性，例如，“女装”指向“箱包皮具”表示购买了“女装”的买家会趋于购买“箱包皮具”。上图中的边都是双向的，应该是所用数据计算得出的结果。
- 在界面右边的“阈值设定”面板，用户可以设定关联规则的最小支持度和最小置信度，只有大于这两个阈值的关联规则才会被显示出来。

- 在界面右边的“属性选择”面板，用户可以选择具有某些特定属性的人群进行关联规则的计算。例如，同时购买女装和女鞋的买家在全部用户中进行统计时，支持度和置信度可能不高，但是如果在女性用户中进行统计，支持度和置信度可能会相当高，因此，有必要让用户交互选择在哪些特定人群中进行关联规则的统计计算。
 - 右下角是“统计信息”面板，用户交互选择某些关联规则，在上图显示的例子中，统计的是同时购买了“女装”、“美容护肤”、“箱包皮具”和“话费充值”的人群信息。包括男女所占比例，年龄所占比例，职业所占比例。
- 4) 几个问题：
- 由于只有 10000 个用户数据，因此我觉得统计出来的关联结果可信度不高。而且，如果是对更细的类目层级进行统计，结果应该更有意义。
 - 对于关联规则的可视化可以再挖掘，例如代表每个类目的形状、布局等。
 - 计算模型的改进，使其更具有通用性。

下周计划：

1. 淘宝的任务估计还要继续完善。
2. 完成肝脏结题工作。
3. 继续标签项目的探索。